VCFN: Virtual Cloth Fitting Try-on Network.

Muhammad Usman Ghani [1], Abdullah Tariq [2], Tanzila Saba[3], Amjad Rehman[3], Hoshang Kolivand[4,5]

[1] Department of Computer Science, Intelligent Criminology Lab, AlKhawarizmi Institute of Computer Science, University of Engineering and Technology, UET Lahore, Pakistan
[2]Al Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore Pakistan
[3]Artificial Intelligence & Data Analytics Lab (AIDA) CCIS Prince Sultan University, Riyadh, 11586 Saudi Arabia
[4]School of Computer Science and Mathematics, Liverpool John Moores University, Liverpool, UK
[5]School of Computing and Digital Technologies, Staffordshire University, Staffordshire, UK

***Abstract*— Virtual cloth fitting network has an increasing demand with a growing online shopping trend to map target clothes on reference subject. Previous research depicts limitations in the generation of promising deformed clothes on the wearer's body while retaining the design features of cloth-like logo, text and wrinkles. The proposed model first learns thin-plate spline transformations to warp images according to body shape, followed by a try-on module. The former model combines deformed cloth with a rendered image to generate composition mask and outputs target body without blurry clothes while preserving critical requirements of the wearer. Experiments are performed on the Zalando dataset and the model produces fine richer details and promised generalized results.**

Keywords— UNet, thin-late spline transformations, cloth deformation, DAMM, VCFN, person representation, technological development

I. INTRODUCTION

Virtual apparel fitting models acquiesce their customers to try clothes without physically wearing them. With an increasing demand for online shopping, the buyer wants to know-how he will look in the desired fashion item when purchasing it online. Thus, with the convenience of online shopping, self-reliance on buying a fashion item would be more comfortable for consumers and enhance their shopping experience. Retailers will also expect reasonable costs in the clothing market. The proposed technique is cheap and rapid compared to photoshoots of each dress on models.

Existing virtual try-on networks are commercially deployed like Fitnect, which provides 3D animation of body movement with desired cloth. However, generating 3D pose movement for bridal or party-wear dresses with fine work is challenging. Over and above that, fabric, for instance, silk velvet, chiffon and organza have an elegant structure which can be lost in 3D modeling of getting dressed. It may include flared sleeves or gowns, animation of which demands complex 3D fashion model approach. This accession is not so accurate in generating real-time movement with a desired cloth. It can also disgrace consumers' usage and may cause a downturn clothing market. 3D modeling is also costly and needs high-priced hardware with immense computation. Furthermore, another approach is to work on 2D image of user for virtual fitting [1]. Wang et al., [2] preserves aforementioned characteristics but visualization can be enhanced by using effective skip connections and an efficient encoder-decoder network. Above all, it is worth saying that 2D implementation of this problem requires less computation and proves to be in more practice by a consumer.

Virtual Cloth Fitting Network (VCFN) is proposed with defined skip connections in Unet generator layers that safeguards the features of input cloth and pose points to cover the gap between retaining features and optimizing the network. An alignment matching module is mainly employed to get warped clothes according to the body pose. For warping purposes, thin plate spline (TPS) transformations are used which fit clothes on target person with the generation of naturally deformed dresses. The proposed learnable model is trained from pairs of human input images and in-shop cloth. VCFN follows a two-stage pipeline for generating target persons. First section of model utilizes person representation suggested in [2] and cloth for processing and outputs deformed clothes aligned with pose of body. In the second section, aligned in-shop clothes and person representation are passed to UNet generator and generate rendered image and composition masks. Furthermore, the composition mask smooths out the wearer's target after working on a pose-coherent image. Significant improvement in this model involves the refinement of Unet generator with specified skip connections, which reduces the loss of VCFN and augments the visualization characteristics of target person.

II. LITERATURE SURVEY

Image generation is performed using Generative Adversarial Networks (GANs) and generating desired images of wearer. Conditional GANs (cGANs) have recently been used for image-to-image translation [3]. Isola et al. [4] proposed edge translation to real image using cGANs and reported improved results with a learning loss function. Wang et al. further upgraded image-to-image translation with appealing visualization using semantic edge map and cGANs [5]. However, unpaired translation of images yields successful results in color or texture, such as in [6] but fails to translate geometric changes of maps. Cross-domain mapping is also dealt with by Kim et al. [7] using DiscoGAN. Their research generated shoes and bags with given cross-domain input and experimented on blond and black but still have limitations in developing toys and fine designs.

Swapping cloth virtually on a person can be visualized in 2D or 3D, as discussed previously in Section I. DRAPE in [8], to fit 2D clothes on 3D body images with aligned poses and body shapes. Hilsmann et al. [9] experimented and proposed a virtual mirror that could replace the texture and color of cloth but restrict users to the same pose and upper-garment specifications like sleeves and style. Han et al., [10] implements virtual try-on module with pose points and encoder-decoder network but fails to retain characteristics of

human and cloth (e.g face, hair, complexion, pose, cloth's design, log and other fine details). However, fail to preserve characteristics of clothes on new target body [11]. In its improvement, Wang et al., [2] aim to retain fine features of clothes on the wearer's body and uses TPS transformations in his work for alignment. For better-generalized deformation and cloth fitting, proposed model enhances the tryon system results after experimenting with different modules with and without skip connections of UNet and reports better visualization to customers at the end [12]. Hashmi et al. [13] proposed the neural fit body method for visualizing the clothes on users by using the GANs to generate the segmented images and NBF for fitting the clothes according to the user's body. In this research, deep fashion2 dataset is used for experiments and attained 97% accuracy. Dong et al. [14] developed a multi-pose virtual try-on network. Warp GAN and conditional human parsing network is implemented for estimating the human pose and outfitting the cloth to human. A new self-collected dataset is used called MPV and also performed experiments on deep fashion dataset. For customarily fashion item tryon, FashionGAN reads a textual representation of desired cloth along with person image and outputs modified image but limits user to clothes without logo and fine designs. Related to this, CAGAN swaps upper-body clothes and trained models on loss function but geometry matching is an excellent concern which is the limitation in his work. We desire to swap clothes without losing person's identity and fine alignment of body pose with clothes.

Further this research has three main sections. Section 3 presents proposed methodology, section 4 exhibits results & analysis. Finally, research is concluded in section 5.

III. Methodology

To address virtual try-on problem, VCFN aims to provide wearer image without losing its face, hair, pose and shape information. Customarily, a person X wants to try cloth c and his ground truth is X_g, VCFN generates wearer image in desired clothes.

One way to synthesize target is to train X directly on X_g in triplet (X, c, X_g) but it has drawbacks. This is straightforward to pass X and c as input to obtain X_g given X and X_g are same to generate results. If any decoupled X and X_g are provided in evaluation, model will not handle wearer's desire. To address this issue, two-way pipeline is followed in [2], in which person representation h is used in the form (h, c, X_g) to train and get generalized results. VCFN demands alignment of clothes with body shape and pose and is a very challenging problem. Existing image generation networks like UNet, ResNet and GANs [4] are deficient in producing deformation clothes and have unclear try-on images. We suggested a deformation and alignment matching module (DAMM) that aligns in-shop clothes with wearer's body to deal with this difficulty. The complete pipeline of networks including DAMM and try-on modules is discussed in next sections.

A. Person Presentation

Person representation h as shown in Figure 1 (b) follows concatenation of three individual representations of a person to generate one vector of size (256×192×22). Detail of vector h is mentioned below.

Human parser: Ground truth human image is segmented with multi-scale segmentation technique to distinguish face, hair, clothes, skirts, bags, pants and shoes etc. Pretrained JPPNET model for LIP dataset proposed by Liang et al. [12] is used to obtain human parsing images. Image parsing employed in [12] uses DeepLab (ResNet-101). These results are generated by using pre-trained weights. The generated segmentation map is further converted to palletized image with one channel binary mask as shown in Figure 1 (a).

Pose Estimator: State-of-the-art pose estimator proposed by Cao et al. [15] is used to extract 18 coordinate points of body in the form of 18-channel feature map to generate wrappings.

Facial and hair segment: Face, complexion and hairstyle of wearer is preserved by extracting masks of these segments from parser map and recovering RGB 3-channel identity regions.

These 3 feature maps are concatenated as one portrayal map with 1+18+3 = 22 channel vector of person preserving its attributes to utilize them to generate the target image. This representation is injected into both DAMM and VCFN modules to propagate processing.

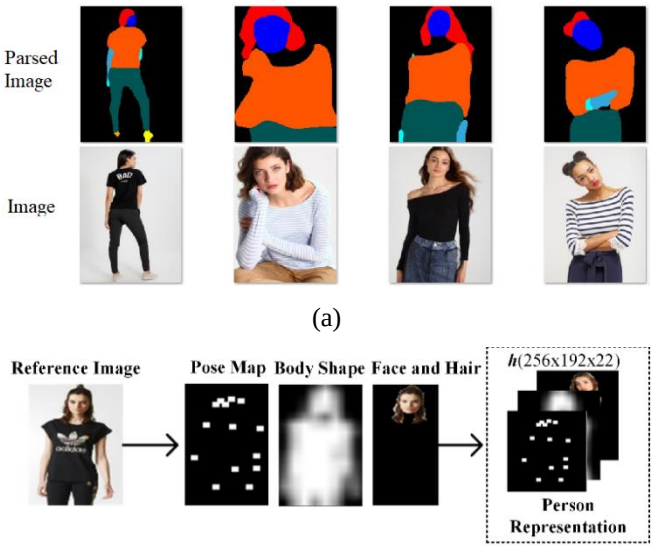

Fig. 1. (a) Concatenation of pose map, body shape, face and hair to produce person representation (b) Given person image and corresponding parsed image

B. Deformation and Alignment Matching Module

To avoid pasting cloth on body, another technique is being employed for warping of clothes. Its main purpose is to align shape of cloth with human body. This deformation and alignment matching module (DAMM) consists of following steps and is exhibited in Figure 2.

1. The model takes person representation h and cloth c, to compute high-level features.
2. Feature maps of h and c are convoluted separately and merged using correlation layer. Finally, it compares vectors with patch size K = 2k + 1 given m_1 and m_2 as feature maps where c_1 and c_2 are centers of h and c respectively as shown in equation 1.

$$\left(cor(c_1, c_2) = \sum_{o \in [-k,k] X [-k,k]} \langle m_1(c_1 + o), m_2(c_2 + o) \rangle \right) \quad (1)$$

3. To transform the parameters for warping single tensor matrix is fed to regression network to foresee transfiguration framework. Spatial transformation is mapping that corroborates the resemblance between input image and warped output. Where p and q are output image coordinates while X an and Y are mapping functions transforming a and b input image coordinates as presented in equation 2.

$$[p, q] = [x(a, b), y(a, b)] \quad (2)$$

4. Headway, Thin Plate Spline (TPS) is an efficient tool for deformation of clothes to transform matrices coordinates. For example, Nystro¨m method [11] proposes that if the input image has i×I pixels, it can be mapped to the target image with deformed points. Concisely, TPS uses transformation module T, outputs contorted and warped cloth c′ = Tθ(c). Take a look inside the module, consider K a matrix involved in yielding TPS coefficients is represented in equation 3.

$$K = \begin{bmatrix} P & Q \\ Q^{-t} & R \end{bmatrix} \quad (3)$$

Where P Rm×m, Q Rm×n and R Rn×n with m and n subsets of total image points i, R can be computed using equation 4.

$$R = Q^t P^{-1} Q \quad (4)$$

DAMM is trained on triplets (h; c; cg) where c is hanged cloth to be warped and cg is ground truth cloth worn on target person. Warping is regulated based on pixel-wise loss L1, comparing predicted vector c′ and ground truth vector cg on each pixel exclusively as demonstrated in equation 5.

$$L_{DAMM}(\theta) = \left|\left| c - c_g \right|\right| = \left|\left| T_\theta(C) - c_g \right|\right| \quad (5)$$

C. Cloth Fitting Module

One way to put the output on target body is to paste it on the body directly, but it doesn't give the desired visualization. It has some advantages like preserving fine warping but looking unnatural at boundary of clothes. Another way to generate controlled fashion-related images is through encoder-decoder network such as UNet used in this paper. UNet generates smooth image but a little misalignment can lead to blurry or misfit cloth on wearer's body. VCFN aims to fit clothes on body by training UNet to synthesize rendered images and composition masks. UNet takes person representation h along with warped cloth c′ and translates it to rendered image Xr with composition mask M. This further leads to fusion of Xr with M and warped cloth c′ to smooth the edges and clothes on body, and form output Xo as presented in equation 6 where matrix multiplication is represented by Equ: 6.

$$X_0 = M \odot c' + (1 - M) \odot X_r \quad (6)$$

We trained the try-on module on triplet (h; c; Xg) to minimize the divergence between Xo and Xg. For this purpose, total loss of network is computed by the sum of L1 loss from equation 7.

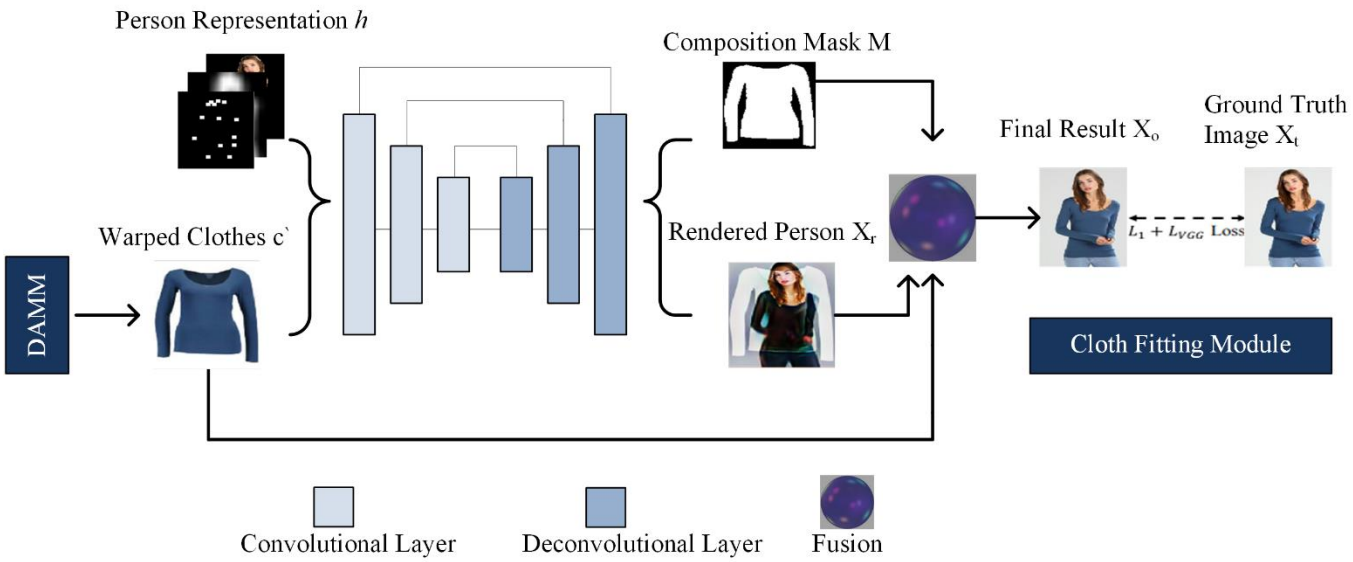

Fig. 2. Spectrograms in form of frequency patterns to visualize STFT and MDCT features of clips after applying audio operations

$$L_i\left(X_0, X_g\right) = \sum_{i=1}^{5} \left| X_0 - X_g \right| \quad (7)$$

and VGG perceptual loss in equation 8.

$$L_{VGG}\left(X_0, X_g\right) = \sum_{i=1}^{5} \left|\phi_i\left(X_0\right) - \phi_i\left(X_g\right)\right| \quad (8)$$

Where φ represents perception network i.e VGG19 pretrained on ImageNet with i layers such as conv1-2, conv2-2, conv3- 2, conv4-2, conv5-2 and φi(X) shows a feature map of X of ith layer. λ restricts the contribution of layers in overall loss. Total network loss is given in equation 9 where M is biased to choose warped clothes with regularizing k1 − Mk1 loss.

$$L_{VCFN} = \lambda_{L_1} L_1\left(X_0, X_g\right) + \lambda_{\text{vgg}} L_{VGG}\left(X_0, X_g\right) + \lambda_{\text{mask}} \left|\left|1 - M\right|\right| \quad (9)$$

IV. EXPERIMENTAL RESULTS & ANALYSIS

A. Dataset

The Zalando dataset [10] was used in 2D try-on experiments. This fashion dataset comprises 16253 pairs of frontal-view women and cloth, out of which 14221 are used for training and experiments. In comparison, proposed model is evaluated and tested on 2032 pairs.

Deformation and Alignment Matching Module: Feature extraction for person representation h and proposed DAMM network for warping contains convolutional layers with filters 64, 128, 256, 512, 512 respectively with first four down sampling layers having stride size 2 and last layers with size 1. Next, regression network with four convolutional layers and one fully connected layer is employed with filters 512, 256, 128, 64 respectively. In regression network, one fully connected layer, two convolutional layers of stride 2 and the following two of stride size 1 predicts TPS points.

Try-on Module: consists of encoder-decoder network having 6 down-sampling and 6 up-sampling layers with stride 2. To lessen checker-board artefacts, the deconvolution layers of UNet are recouped with nearest-neighbour interpolation layers.

B. Quantitative Comparison

To validate the results of VCFN, proposed model is compared with previous baselines and quantitative scores are analyzed. CAGAN synthesizes the target image but needs an original cloth image and a reference image at test time, making it impractical for consumers. On the other hand, CRN has the advantage of generating high-resolution images due to the usage of refinement networks. Moreover, VITON [10] produces target person using context matching module and refinement network but fails in preserving characteristics of cloth. In most recent work, CPVTON [2] synthesizes wearer's body with improvements in preserving pose of body, style and logo of clothes.

TABLE I: Comparison of the proposed approach with a baseline modules

Method /ref	IS (Inception Score)	Human	Time in Sec
JPPNet[12]	2.980+0.167	27.6	120
CRN[15]	2.447+0.170	69.5	132
VITON[6]	2.514+0.110	77.3	116
CP-VTON [4]	2.589+0.130	84.1	97
without DAMM	2.700+0.210	68.6	93
without pose estimation	2.291+0.140	66.7	84
without skip connections	2.600+0.110	81.5	98
with skip connections (VCFN)	2.683+0.130	88.8	80

A quantitative measure of this try-on w/o DAMM is presented in Table I. The IS is the inception score measure to quantify image quality [10]. Some of the sample results of proposed method are shown in Figure 3. Furthermore, person representation excluding pose points is also considered and tested for try-on purpose but yields failed results with poor alignment. Row with label w/o pose estimation in Table I shows evaluation score on this module. Try-on module experimented with UNet generator network without skip connections shown in Table I reduces the model's effectiveness and increases computation time by 10s. Beyond these attempts, try-on module with person representation, DAMM and UNet with skip connections is assessed and improved results are observed. The quality of target persons generated in previously discussed baseline models and proposed models can be evaluated using inception score (IS). IS is used to estimate the semantics and credibility of synthesized images. Images with high reliability will have high inception scores. Column 1 in Table I shows IS for discussed modules, whereas column 2 measures absoluteness and reality of generated human images.

Fig. 3. Sample results from proposed VCFN

For verification of pose and reference body Amazon Mechanical Turk (AMT) and human score measure whether the synthesized body, cloth and face are real or not, as stated in [10] is employed. For example, for target wearer images from proposed VCFN with skip connections, AMT shows 86% of images as real.

V. CONCLUSION AND FUTURE WORK

In this work, we contemplated a try-on module following a two-way structured pipeline that transforms the in-shop clothes according to body shape and then transfers garments onto the wearer's body. TPS transformations efficiently deform the cloth in the DAMM module and generate warped clothes after passing through UNet, skip connections produce a composition mask. The rendered image gets smoothened with mask and provides the customer an excellent visual

representation of the target. Extensive experiments are performed to achieve promising results. This work concluded that 2D images yield customer-friendly visualizations with less computation. In the future, we plan to add a wide range of fabrics in VCFN, preserving the quality and grace of clothes. We will further work on accessories like bags and necklaces to make the system more desirable for customers.

ACKNOWLEDGMENT

This work was done in collaboration with Seventh Reality pvt ltd [http://seventhreality.com/] a software company specialized in development of AI specific applications related to video surveillance and mixed reality domains. This research is also supported by Artificial Intelligence & Data Analytics Lab CCIS Prince Sultan University Riyadh Saudi Arabia.

REFERENCES

[1] Basori, A. H., Alkawaz, M. H., Saba, T., & Rehman, A. (2018). An overview of interactive wet cloth simulation in virtual reality and serious games. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 6(1), 93-100.

[2] Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., & Yang, M. (2018). Toward characteristic-preserving image-based virtual try-on network. In Proceedings of the European conference on computer vision (ECCV) (pp. 589-604).

[3] Dosovitskiy, A., Tobias Springenberg, J., & Brox, T. (2015). Learning to generate chairs with convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1538-1546).

[4] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).

[5] Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8798-8807).

[6] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

[7] Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017, July). Learning to discover cross-domain relations with generative adversarial networks. In International conference on machine learning (pp. 1857-1865). PMLR.

[8] Guan, P., Reiss, L., Hirshberg, D. A., Weiss, A., & Black, M. J. (2012). Drape: Dressing any person. ACM Transactions on Graphics (ToG), 31(4), 1-10.

[9] Hilsmann, A., & Eisert, P. (2009, May). Tracking and retexturing cloth for real-time virtual clothing applications. In International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications (pp. 94-105). Springer, Berlin, Heidelberg.

[10] Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L. S. (2018). Viton: An image-based virtual try-on network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7543-7552).

[11] Donato, G., & Belongie, S. (2002, May). Approximate thin plate spline mappings. In European conference on computer vision (pp. 21-31). Springer, Berlin, Heidelberg.

[12] Liang, X., Gong, K., Shen, X., & Lin, L. (2018). Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE transactions on pattern analysis and machine intelligence, 41(4), 871-885.

[13] Hashmi, M. F., Ashish, B. K. K., Keskar, A. G., Bokde, N. D., & Geem, Z. W. (2020). FashionFit: Analysis of mapping 3D pose and neural body fit for custom virtual try-on. IEEE Access, 8, 91603-91615.

[14] Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., ... & Yin, J. (2019). Towards multi-pose guided virtual try-on network. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9026-9035).

[15] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).